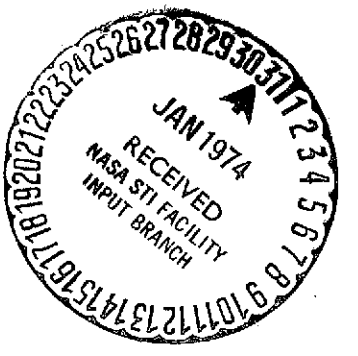
1 up

NASA TECHNICAL TRANSLATION NASA TT F-15274

FINDING THE MOST CONSISTENT MOLECULAR WEIGHT BY ANALYSIS OF THE AMINO ACIDS OF A PROTEIN

M. Delaage

Translation of "Sur la recherche du poids moléculaire le plus cohérent avec l'analyse des acides aminés d'une protéine," Biochim. Biophys. Acta, 168, 1968, pp. 573-575.



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546 JANUARY 1974

STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
TT F-15274		

4. Title and Subtitle	5. Report Date
FINDING THE MOST CONSISTENT MOLECULAR WEIGHT BY ANALYSIS OF THE AMINO ACIDS OF A PROTEIN	January 1974
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
M. Delaage Centre de Biochimie et Biologie Moléculaire C.N.R.S., Marseille (France)	10. Work Unit No.
9. Performing Organization Name and Address	11. Contract or Grant No.
Leo Kanner Associates Redwood City, CA	NASw 2481
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Translation
National Aeronautics and Space Administration Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

Translation of "Sur la recherche du poids moléculaire le plus cohérent avec l'analyse des acides aminés d'une protéine," Biochim. Biophys. Acta, 168, 1968, pp. 573-575.

16. Abstract

A method is given for determining the molecular weight of a protein on the basis of its amino acid composition. Two procedures are outlined, one using a computer and the other mathematical tables. Possibilities for error are discussed briefly.

17. Key Words (Selected by Author(s))	18. Distribution Statement
	Unclassified - Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	8	3.00

FINDING THE MOST CONSISTENT MOLECULAR WEIGHT BY ANALYSIS OF THE AMINO ACIDS OF A PROTEIN

M. Delaage

Today analysis of the amino acid composition of proteins has reached a sufficient degree of accuracy that it may be used as a method for determining molecular weight.

I. Principles of Computation

Analysis of a protein sample has given, once all corrections have been taken into account, the number of μmoles x_i of each amino acid (i). The unknowns are x, the number of μmoles of protein analyzed, and the N_i values, that is the number of amino acid residues of type (i) in a mole of protein. If the analysis is correct the result should be:

$$x_i = N_i x \tag{1}$$

All that remains is to find x, the largest common denominator for the x_i values, and subtract the N_i values, and the molecular weight is obtained.

If the analysis has not been accurate, Eq. (1) will not be verified. It is known that, for the actual value of x, the aleatory variable $n_i = x_i/x$ is distributed normally around N_i and furthermore that the reduced variable n_i/N_i also follows a normal distribution no matter what amino acid (i) is involved. As a result the quantity Y:

$$Y = \sum_i \left(\frac{n_i}{N_i} - 1\right)^2$$

behaves as a random χ^2 variable with a number of degrees of freedom equal to the number of amino acids. If x is fixed arbitrarily and N_i is given the integer value closest to the ratio $x_i/x = n_i$, the Y function measures the difference between the composition obtained and the nearest integral composition.

The value for x which should be used is that which places Y at a minimum. x is directly linked to the molecular weight P. In effect the weight p of the sample is precisely known:

$$p = \sum_i x_i p_i \qquad (2)$$

p_i: molecular weight of residue of amino acids (i). The molecular weight associated with x is $P = p/x$. Either the curve $Y(P)$ or the curve $Y(x)$ can be constructed.

Plotting the function $Y(P)$ should reveal a pronounced minimum for the actual value of the molecular weight and the multiples of this value. There may be a local minimum each time the molecular weight is in simple ratio to the actual molecular weight: 1/2, 3/2, etc.

2. Computation Technique

(a) On computer. Computation is direct. The computer is given the x_i values, the interval of variation in x corresponding to the extreme values for P, for example 10,000-30,000, and the step of the variation; the difference between two consecutive values should not exceed 2%. For each x value the computer calculates the $x_i/x = n_i$ ratios, finds the nearest integer value for N_i, determines Y, changes the value for x, and begins the process over again.

(b) Without computer. A different procedure is preferable. When the differences are small (as is generally the case), the

following approximation applies:

$$\left(\frac{n_i}{N_i} - 1\right)^2 \simeq \left(\mathrm{Log}\frac{n_i}{N_i}\right)^2$$

Log = neper log. Y can therefore be replaced by a function y which is approximately proportional to it:

$$y = \sum_i \left(\log \frac{n_i}{N_i}\right)^2$$

log = decimal log. For each log x value, log n_i = log x_i - log x is calculated, a log table for numbers from 1 to 100 is used to locate the nearest log to log n_i, that is, log N_i, and the difference between log n_i - log N_i is found; this is squared using a table of squares for numbers from 0 to 100. The values obtained are added to obtain y.

The logs will be adequate taken to three decimal places; to proceed from one value for log x to the next a step size of 0.010 is selected and the new values for log n_i can be calculated from the former ones.

Fig. 1 shows a theoretical plotting of the function Y in the neighborhood of its principal minimum, as would be obtained with a perfect analysis of chymotrypsinogen A from cattle in accordance with the sequences constructed by Hartley et al. [1] or Sorm [2]. (Glutamine and asparagine are included along with their corresponding amino acids.)

Fig. 2 shows the y function actually obtained in an experimental analysis of chymotrypsinogen B from cattle by Gratecos et al. [3]. The molecular weight obtained, 25,500, is in excellent agreement with that of 25,754 resulting from the sequence constructed by Smillie et al. [4], as well as with the ultracentrifugation data: 25,500 $\pm$ 1500 [6].

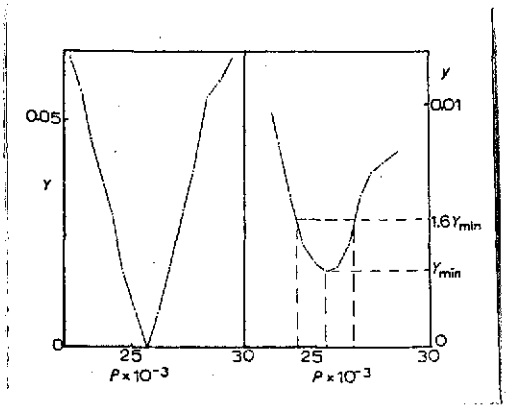


Fig. 1 (left). Theoretical curve for

for chymotrypsinogen A from cattle. Computation made on the Olivetti Programma 101.

Fig. 2 (right). Experimental curve for

for chymotrypsinogen B from cattle [3]. Computation made using tables. The molecular weight corresponding to the minimum has a surrounding confidence factor of 5%. The ordinates are directly comparable to those in Fig. 1, the scale being $(\log e)^2 = 0.198$.

The method can also be applied to large-size proteins: for the A type monomer of the alkaline phosphatase of *Escherichia coli* identified by Lazdunski and Lazdunski [6], the minimum for Y is obtained at a molecular weight of 47,200, that is, 94,400 for the dimer; the molecular weight found by ultracentrifugation is 86,000 [7].

Once the configuration for Y has been obtained and the minimum corresponding to the actual molecular weight has been precisely determined, several points may be questioned:

(a) The accuracy of the result obtained. If the minimum difference observed Y_{min} corresponds to the median of the Pearson distribution at 20 degrees of freedom, the probability that a difference of $1.6\ Y_{min}$ will be exceeded is only 0.05. The confidence factor surrounding the most likely molecular weight can be estimated at the total number of neighboring points for which: $Y \leq 1.6\ Y_{min}$: for example, $P = 25{,}500 \pm 1200$ in Fig. 2, but it would be easy to reach $\pm$ 500.

(b) Computation of the n_i values at minimum Y theoretically leads to a group of values which are very close to the integer values N_i. One of these n_i values, however, may diverge widely from the N_i values, thus placing the results of the analysis in doubt for the amino acid under consideration.

Without destroying the accuracy of the calculations a particular amino acid may be left out of consideration, and more generally, if accurate data on the individual standard differences of amino acids are available they may be used to weight the contributions of these amino acids in the Y expression. At the end of computation two factors will be added to the molecular weight: the value for the residues set aside, estimated as accurately as possible, and one molecule of water by open chain.

In conclusion, although this method requires a high-purity protein, on the other hand it requires only a small quantity of the substance. Theoretically it permits independent determination of the unit molecular weight, and when the general range for this value has already been found by physical methods it should permit a closer approximation of the actual value.

I would like to take this opportunity to thank Ms. D. Gratecos and Mr. C. Lazdunski and Professors P. Desnuelle and M. Lazdunski for the interest they have taken in this work and for their many helpful discussions with me.

REFERENCES

1. Hartley, B.S., Brown, J.R., Kauffman, D.L., and Smillie, L.B., Nature, 207, 1157 (1965).

2. Sorm, F., Holeysovsky, V., Mikes, O., and Tomasek, V., Collection Czech. Chem. Commun., 30, 2103 (1965).

3. Gratecos, D., Guy, O., Rovery, M. and Desnuelle, P., Biochim. Biophys. Acta, publication currently under way.

4. Smillie, L.B., Furka, A., Nagabhushan, N., Stevenson, K.J., and Parkes, C.O., Nature 218, 343 (1968).

5. Guy, O., Gratecos, D., Rovery, M., and Desnuelle, P., Biochim. Biophys. Acta, 147, 280 (1967).

6. Lazdunski, C., and Lazdunski, M., Biochim. Biophys. Acta, 147, 280 (1967).

7. Schlesinger, M.J., and Barrett, K., J. Biol. Chem. 240, 4284 (1965).